1E

УДК 004.934

Т.В. Ермоленко, А.В. Лащенко
Институт проблем искусственного интеллекта МОН и НАН Украины, г. Донецк, Украина
etv@iai.donetsk.ua

Применение вейвлет-анализа для определения границ речи в зашумленном сигнале

В статье предложена методика определения границ речи в звуковом сигнале, содержащем шум, на основе вейвлет-анализа. Одним из этапов этой процедуры является классификация фреймов входного сигнала, основанная на энергетических характеристиках вейвлет-спектра и позволяющая учитывать акустические характеристики широких фонетических классов звуков речи. Подобный подход дает возможность определить границы речи при наличии высокоамплитудных помех, провести сегментацию речевого сигнала и повысить эффективность дальнейшего распознавания.

Введение

Одним из важных направлений исследований в области искусственного интеллекта является разработка интеллектуальных систем образного восприятия речевой информации, среди которых значительную роль играют системы распознавания речи. Проблемы, возникающие при распознавании речевого сигнала, связаны с его вариативностью, шумом окружающей среды и звукозаписывающего оборудования, поэтому качество распознавания существенно зависит от предварительной обработки сигнала.

Одним из этапов предварительной обработки речевого сигнала является определение границ речи. Соответствующие методы реализованы в многочисленных детекторах речи (VAD). Общим свойством VAD-алгоритмов является то, что они включают в себя обучение (вычисление характеристик шума) и спектральное вычитание. Чаще всего в качестве признаков, определяющих начало и конец слова, выбираются энергетические и спектральные характеристики сигнала [1-3], а также число переходов через ноль [4], [5]. К недостаткам VAD-алгоритмов, базирующихся на оценке энергетических характеристик сигнала, относится возможность принятия кратковременного шума с высокой амплитудой за речь либо низкоамплитудного речевого сигнала за шум. Корректно работающий в подобных ситуациях детектор описан в [6], в качестве признаков классификации речь/шум используются мел-частотные кепстральные коэффициенты. Однако для его эффективной работы необходимо наличие в обучающем множестве как сигнала, содержащего только шум, так и речевых баз данных.

Кроме того, большинство из VAD не способны точно определять границы речи в условиях шума, уровень которого превышает или близок к уровню шумных глухих щелевых и смычно-щелевых звуков. Для решения этой проблемы необходимо при формировании набора признаков, определяющих начало и конец слова, учитывать спектральные характеристики широких фонетических классов (ШФК) звуков речи, а также их длительность.

Для описания локальных особенностей неоднородных сигналов, к которым относится речевой сигнал, в последнее время эффективно употребляется вейвлет-преобразование, которое обеспечивает подвижное частотно-временное окно анализа и адаптировано к локальным свойствам сигнала [7], [8].

В данной работе на основе вейвлет-анализа предлагается методика определения границ речи в звуковом сигнале, позволяющая выделить речь при наличии высокоамплитудных помех за счет учета акустических характеристик ШФК звуков речи с одновременной первичной сегментацией речевого сигнала.

Под термином «первичная сегментация» в данной работе понимается разбиение сигнала на участки, каждый из которых содержит один из следующих ШФК звуков речи:

– шум (*Noise*);
– вокализованный звук (*Voc*);
– шумный глухой щелевой или смычно-щелевой звук (*Sh*);
– шумный глухой смычный звук (*P*).

Методика определения границ речи

Предложенная ниже методика определения границ речи использует быстрое вейвлет-преобразование Добеши [9] и состоит из трех этапов: обучения шуму, классификации фреймов сигнала, определения границ речи (рис. 1).

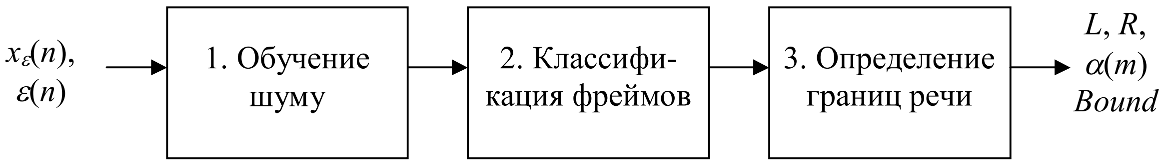


Рисунок 1 – Схема методики определения границ речи

Входными данными этой процедуры являются зашумленный сигнал $x_\varepsilon(n)$ и образец шума $\varepsilon(n)$; выходными данными – отсчеты сигнала L, R, которые соответствуют левой и правой границам слова, вычисленные по образцу шума на каждом уровне разложения; пороги $\alpha(m)$ и массив номеров граничных фреймов, полученный в результате классификации фреймов.

На этапе обучения шуму выполняется вейвлет-разложение сигнала $\varepsilon(n)$, его разбиение на фреймы длиной ΔN, образующие множество фреймов F_ε, и вычисление порогов $\alpha(m)$:

$$\alpha(m) = Aver_\varepsilon(m) + 3\sqrt{D_\varepsilon(m)}\,,\ m = 1,\ldots,j_{\max}, \tag{1}$$

где $j_{\max}$ – максимальный уровень вейвлет-разложения; $Aver_\varepsilon(m)$, $D_\varepsilon(m)$ – полученные на множестве F_ε среднее и смещенная оценка дисперсии величин (2), представляющих собой энергии спектра $E_s^\varepsilon(m)$ сигнала $\varepsilon(n)$

$$E_s^\varepsilon(m) = \sum_{n=(s-1)\Delta N/2^m}^{s\Delta N/2^m} d_{mn}^2\,,\ \mathrm{s} \in F_\varepsilon\,. \tag{2}$$

На этапе классификации каждый фрейм входного сигнала $x_\varepsilon(n)$ относят к одному из четырех ШФК, перечисленных выше. Классификация фреймов проводится на множествах уровней разложения: $M_{voc}=\{m\colon m_{voc} \le m \le j_{\max}\}$ – соответствует полосе частот основного тона (100 – 300 Гц); $M_{sh}=\{m\colon 1 \le m \le m_{sh}\}$ – соответствует высокочастотной области спектра (более 2500 Гц), где сосредоточена энергия звуков класса *Sh*.

Рис. 2 демонстрирует поведение характеристик (2) для сигнала, записанного в условиях высокоамплитудного производственного шума (отношение сигнал/шум 2,3 дБ), содержащего звуки различных ШФК (рис. 2а), на уровне $m \in M_{sh}$ (рис. 2б) и $m \in M_{voc}$ (рис. 2в).

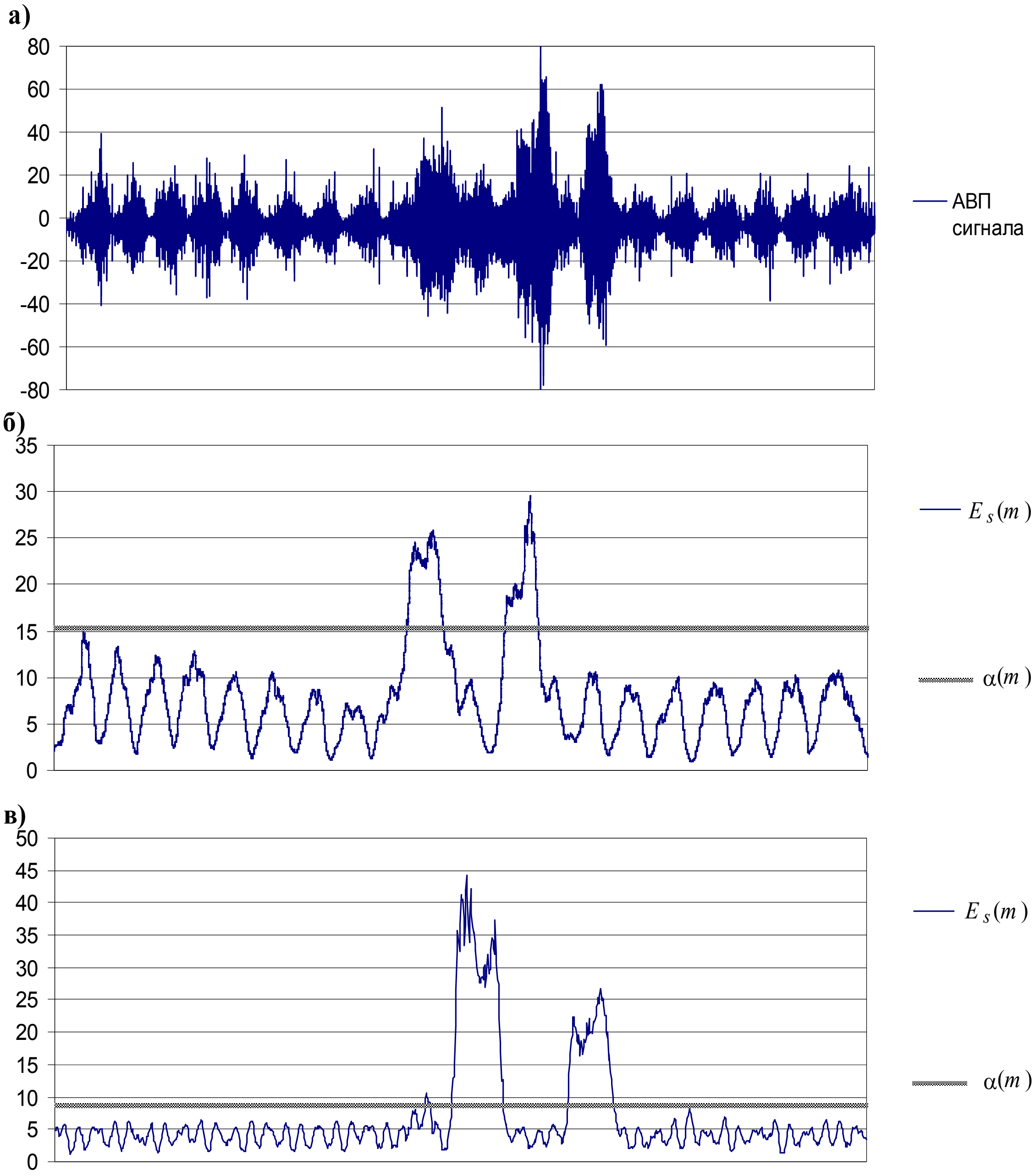


Рисунок 2 – а) Амплитудно-временное представление слова «Сушка», записанного в условиях производственного шума, б) энергия вейвлет-спектра слова «Сушка» на уровне $m \in M_{sh}$; в) энергия вейвлет-спектра слова «Сушка» на уровне $m \in M_{voc}$

Как видно из рис. 2, амплитудно-частотные характеристики банка вейвлет-фильтров позволяют на множестве уровней разложения M_{voc} выделить из сигнала вокализованные звуки, на множестве уровней разложения M_{sh} – звуки класса Sh.

Классификация фреймов сигнала проводится по следующим правилам:

$$E_s(m) \le \sum_{m \in M_{Voc} \cup M_{sh}} \alpha(m) \to s \in Noise \vee P,$$

$$E_s(m) \ge \sum_{m \in M_{Voc}} \alpha(m) \to s \in Voc,$$

$$\left(E_s(m) \le \sum_{m \in M_{Voc}} \alpha(m) \right) \wedge \left(E_s(m) \ge \sum_{m \in M_{sh}} \alpha(m) \right) \to s \in Sh,$$

где $E_s(m)$ – энергия s-го фрейма сигнала $x_\varepsilon(n)$.

На основе классификации фреймов строится функция их маркировки:

$$Mark(s) = \begin{cases} 0, & s \in Noise \vee P \\ 1, & s \in Voc \\ 2, & s \in Sh \end{cases}. \qquad (3)$$

Чтобы не принимать кратковременный высокоамплитудный шум за речь, необходимо уточнить маркировку фреймов с учетом минимальной длительности фонемы согласно правилу:

$$\exists N1, N2: (0 \le N2 - N1 < L_{\min}) \wedge (Mark(N1) = Mark(N2) = 0) \wedge$$
$$(Mark(N1 + 1) \ne 0) \wedge (Mark(N2 - 1) \ne 0) \to \forall s{:}N1 \le s \le N2 \quad Mark(s) = 0,$$

где $L_{\min}$ – число фреймов, соответствующее максимальной длительности фонемы.

Следующий этап – определение границ речи. Номера отсчетов L и R, которые являются левой и правой границами речи, определяются согласно (4) и (5):

$$\exists N_l: (\forall s < N_l\ Mark(s) = 0) \wedge Mark(N_l) \ne 0 \to L = N_l \Delta N, \qquad (4)$$

$$\exists N_r: (\forall s: N_r < s \le N_r + L_{\max}\ \ Mark(s) = 0) \wedge Mark(N_r) \ne 0 \to R = N_r \Delta N, \qquad (5)$$

где $L_{\max}$ – число фреймов, соответствующее максимальной длительности звука класса P; ΔN – длина фрейма; N_l, N_r – номера фреймов, соответствующих левой и правой границам речи.

Чтобы не принимать низкоамплитудный речевой сигнал за шум, уточняется маркировка фреймов следующим образом:

$$\forall s: (N_l < s < N_r) \wedge (Mark(s) = 0) \to Mark(s) = 3. \qquad (6)$$

Таким образом, с учетом (6) функция маркировки (3) примет вид:

$$Mark(s) = \begin{cases} 0 & s \in Noise \\ 1 & s \in Voc \\ 2 & s \in Sh \\ 3 & s \in P \end{cases}. \qquad (7)$$

Функция (7) позволяет провести первичную сегментацию речевого сигнала с одновременной классификацией сегментов. Номера граничных фреймов образуют массив (8):

$$Bound = \{s: (N_l + L_{\min} \le s \le N_r - L_{\min}) \wedge (Mark(s - 1) \ne Mark(s))\}. \qquad (8)$$

Результаты численного исследования

Предложенная методика была реализована в виде программного модуля, который является составной частью информационной технологии, реализующей функции предварительной обработки, сегментации речевого сигнала, классификации и распознавания фонем. Тестирование этого модуля проводилось на сигналах, зашумленных цветными шумами, а также содержащих производственные шумы от работающих технических устройств.

В численном исследовании участвовало 50 дикторов с различными голосовыми данными. Каждый диктор произносил слова, содержащие звуки различных ШФК. Слова записывались с частотой дискретизации 22050 Гц, 8 бит, моно. Результаты исследования для сигналов с различными видами шумов сведены в табл. 1, куда при определении границ речи (столбец *Noise*) и сегментов, содержащих звуки классов *Voc*, *Sh*, *P* (столбцы *Voc*, *Sh*, *P* соответственно), занесены: вероятности ошибочного определения границ (столбцы α – вероятность ошибки первого рода) и пропуска границ (столбцы β – вероятность ошибки второго рода).

Таблица 1 – Вероятности ошибок первого и второго рода при определении границ речи и первичной сегментации

Тип шума	*Voc*		*Sh*		*P*		*Noise*	
	α	β	α	β	α	β	α	β
Коричневый шум, отношение сигнал/шум 9 дБ	0,020	0,018	0,022	0,019	0,021	0,019	0,021	0,019
Розовый шум, отношение сигнал/шум 15 дБ	0,045	0,043	0,049	0,030	0,043	0,029	0,049	0,030
Белый шум, отношение сигнал/шум 18 дБ	0,025	0,021	0,041	0,036	0,018	0,015	0,041	0,036
Производственный шум, отношение сигнал/шум 2 – 5 дБ	0,020	0,019	0,024	0,015	0,016	0,014	0,024	0,019

Как можно видеть из табл. 1, вероятности ошибок определения границ речи и сегментов, содержащих звуки разных ШФК, для зашумленных сигналов различными видами шумов не превышают 0,05.

Выводы

Основным результатом данной статьи, отражающим научную новизну, является то, что усовершенствованы методики определения границ речи на основе методов вейвлет-анализа за счет использования акустических характеристик звуков речи, принадлежащих различным ШФК, что дает возможность: определить границы речи в звуковом сигнале при высокоамплитудных помехах, а также в условиях шума, уровень которого превышает или близок к уровню шумных глухих щелевых и смычно-щелевых звуков; провести первичную сегментацию речевого сигнала с одновременной классификацией полученных сегментов. Подобный подход на этапе предварительной обработки позволяет понизить ошибки дальнейшего распознавания.

Численные исследования показали эффективность применения предложенной методики для сигналов, содержащих шумы различных видов, вероятности ошибок при определении границ речи и сегментов не превышают 0,05.

Предложенный подход определения границ речи может быть использован для построения интеллектуальных систем взаимодействия пользователя и компьютера, а также систем речевого управления техническими устройствами.

Литература

1. Аграновский А.В., Зулкарнеев М.Ю., Леднов Д.А., Репалов С.А. Организация иерархической модели распознавания слитной речи // Искусственный интеллект. – 2001. – № 3. – С. 17-22.
2. Freeman D., Sonthcott C., Boyd I.A. Voice activity detector for the Pan-European digital cellular mobile telephone service // IEEE Colloquium «Digitized Speech Communication via Mobile Radio». – London (Great Britain). – 1988. – P. 61-65.
3. Junqua J.C., Mak B., Reaves B. A Robust Algorithm for Word Boundary Detection in the Presence of Noise // IEEE Transactions on Speech Audio Processing. – 1994. – Vol. 2, № 3. – P. 406-412.
4. Редди Д.Р. Машинное распознавание речи // ТИИЭР. – 1976. – Т. 64, № 4. – С. 95-127.
5. Savoji M.H. A Robust Algorithm for Accurate Endpointing of Speech // Speech Communication. – 1989. – Vol. 8, № 3. – P. 45-60.
6. Tomi Kinnunen, Evgenia Chernenko, Marko Tuononen, Pasi Franti, Haizhou Li / Voice Activity Detection Using MFCC Features and SuPort Vector Machine // Proc. International Conf. on Speech and Computer (SPECOM'2007). – Moscow (Russia). – 2007. – P. 556-561.
7. Малла С. Вейвлеты в обработке сигналов: Пер. с англ. – М.: Мир, 2005. – 671 с.
8. Воробьев В.И., Грибунин В.Г. Теория и практика вейвлет-преобразования. – СПб.: ВУС, 1999. – 208 с.
9. Добеши И. Десять лекций по вейвлетам: Пер. с англ. – Москва; Ижевск: РХД, 2004. – 464 с.

Т.В. Єрмоленко, А.В. Лащенко

Методика визначення границь мовлення у сигналі, який містить шум, на базі вейвлет-аналіза

Запропоновано методику визначення границь мовлення у звуковому сигналі, який містить шум, на базі вейвлет-аналізу. Одним із етапів цієї процедури є класифікація фреймів вхідного сигналу, який базується на енергетичних характеристиках вейвлет-спектра та дозволяєураховувати акустичні характеристики широких фонетичних класів звуків мовлення. Такий підхід забезпечує визначення границь мовлення при наявності високоамплітудних завад, надає можливість виконати сегментацію мовного сигналу та підвищити ефективність подальшого розпізнавання.

T.V. Yermolenko, A.V. Laschenko

Wavelet-Analysis Application for Speech Boundaries Detection in a Noised Signal

Wavelet-analysis based method for speech boundaries detection in a noised signal was offered. As one of stages this method includes input signal's frames classification, which is based on wavelet spectrum energy characteristics. It allows to take into account acoustic characteristics of speech sounds' wide classification. Such an approach gives an opportunity to allocate a speech in a signal with high-amplitude noises, to execute a speech signal segmentation and to raise efficiency of further recognition.

Статья поступила в редакцию 16.07.2008.